Dense Multitask Learning to Reconfigure Comics

Deblina Bhattacharjee, Sabine Süsstrunk and Mathieu Salzmann
School of Computer and Communication Sciences, EPFL, Switzerland
{deblina.bhattacharjee, sabine.susstrunk, mathieu.salzmann}@epfl.ch

Abstract

In this paper, we develop a MultiTask Learning (MTL) model to achieve dense predictions for comics panels to, in turn, facilitate the transfer of comics from one publication channel to another by assisting authors in the task of reconfiguring their narratives. Our MTL method can successfully identify the semantic units as well as the embedded notion of 3D in comics panels. This is a significantly challenging problem because comics comprise disparate artistic styles, illustrations, layouts, and object scales that depend on the author's creative process. Typically, dense image-based prediction techniques require a large corpus of data. Finding an automated solution for dense prediction in the comics domain, therefore, becomes more difficult with the lack of ground-truth dense annotations for the comics images. To address these challenges, we develop the following solutions- we leverage a commonly-used strategy known as unsupervised image-to-image translation, which allows us to utilize a large corpus of real-world annotations; - we utilize the results of the translations to develop our multitasking approach that is based on a vision transformer backbone and a domain transferable attention module; -we study the feasibility of integrating our MTL dense-prediction method with an existing retargeting method, thereby reconfiguring comics.

1. Introduction

Comics represent an important part of cultural heritage, preserving decades of artistic expression, stories, societal views, and lore, that predate digital media. Appreciated across age groups, the medium of comics has undergone significant evolution over the past decade. Particularly, there has been a rapid proliferation of digital comics as a consequence of reduced costs, easier transportation, and ubiquitous access. The increasing demands for comics digitization across platforms such as computers, tablets, and mobile phones, call for the automated extraction and identification of relevant elements within comics books. This process of automatically transferring the semantic or graphical units of comics from one publication medium to another is explained as *reconfiguration* of comics. To achieve such reconfigurations across various publishing platforms, a viable step is investigating the relations between the comics elements by means of computational modeling. However, this is a significantly challenging problem due to the 1) disparate styles of comics panels, 2) different text layouts, 3) changing appearances of comics characters, 4) different image scales of elements, panels, etc. Typically, the existing studies that investigate the various elements in comics are restricted to speech balloon segmentation [1,20,33], text detection [32], panel detection [31, 34, 46], comics character detection [10, 26], and region of interest detection [29]. Recently, [3], presented a depth estimation method on comics by exploiting the scene context. While these methods

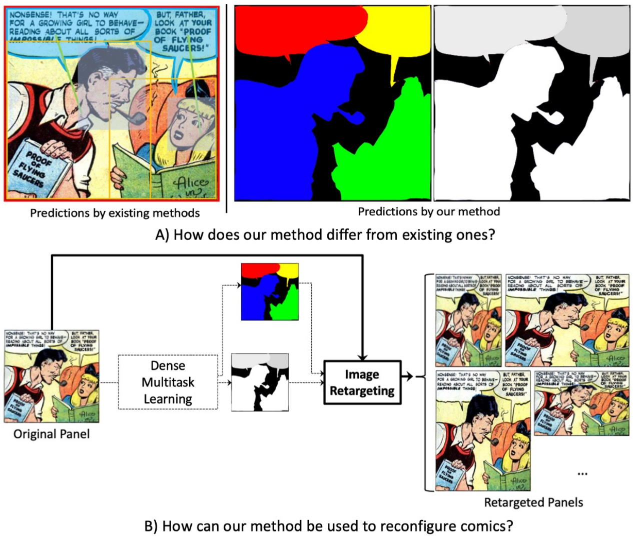
NONSENSE! THAT'S NO WAY FOR A GROWING GIRL TO BEHAVE— READING ABOUT ALL SORTS OF IMPOSSIBLE THINGS!
BUT, FATHER, LOOK AT YOUR BOOK "PROOF OF FLYING SAUCERS!"
PROOF OF FLYING SAUCERS
Alice
Predictions by existing methods
Predictions by our method
A) How does our method differ from existing ones?
Original Panel
Dense Multitask Learning
Image Retargeting
Retargeted Panels
...
B) How can our method be used to reconfigure comics?

Figure 1. **Motivation for our work.** (Top) While existing methods in comics analysis can detect panels (red rectangle), speech balloons (cyan mask), face (light blue rectangular mask), and person (yellow rectangle), we achieve segmentation and depth predictions in a unified manner. (Bottom) Unlike the existing methods, our dense image-based predictions can retarget comics panels, such that it can benefit comics artists to reconfigure their work from one publication medium to another.

achieve promising results, they do not produce strong cues for reconfiguring the comics. Further, none of these existing techniques present a unified approach to produce multiple dense predictions of the graphical elements such as semantic segmentation and depth estimation, simultaneously. In this paper, we present a multitasking method, to segment and estimate the depth of the graphical contents within a comics panel, which are significant cues for reconfiguring comics across different media, as shown in Figure 1. This would help comics authors to diffuse their work across diverse publication channels, thereby benefiting the comics industry. Our contributions are as follows: 1) We introduce a cross-domain multitasking method to perform dense predictions by leveraging an off-the-shelf unsupervised I2I translation method and a vision transformer backbone. 2) We exploit the long-range transformer attention [21] to achieve segmentation and depth predictions in the comics domain. To this end, we use a domain transferable attention mechanism that enforces similarity between the domains. 3) We utilize our dense MTL predictions with an existing retargeting algorithm that successfully reconfigures comics panels across different media.

2. Related Work

Background on Comics Analysis The image analysis community has investigated comic book element extraction for almost 10 years, and methods vary from low-level analysis such as text recognition [1] to high-level analysis such as style recognition [9]. In particular, [27] introduced a deep learning approach to recognize text within the panels and speech balloons by first segmenting the text areas. Differently, [10] processed the graphical units- specifically, the comics characters by leveraging a deep learning-based detection model. Processing such graphical elements is a challenging task because of the ever-evolving appearance of the comics characters, not only across comics books but also across various panels. Nonetheless, several methods have been proposed for recognizing comic characters or faces that encompass deep neural network approaches [10,26,29] or handcrafted feature processing techniques [16, 23, 41]. However, tasks like text recognition, style recognition, or character recognition do not successfully address the challenge of reconfiguring comics as they do *not* produce sufficient dense pixel cues for reconfiguration. Moreover, all the existing methods- based on deep learning approaches or conventional image processing techniques- treat each element within a panel separately. In contrast, our MTL approach can achieve *dense* cues that can accurately reconfigure comics images to different media automatically.

In a different vein, extracting comics panels has been extensively studied [31, 34, 46] to meet the increasing demands of matching the sizes of panels to that of the constantly evolving tablets, portable readers, and smartphones. While matching panel sizes can benefit the reconfiguration of comics panels, it leads to shrinking or stretching artifacts as per the device on which it is reconfigured. This means extracting panels to fit the target device sizes, does not preserve the contents or illustrations as originally intended by the authors of the comics. We mitigate this issue by identifying the key graphical elements within a panel by inferring the semantics and depth of the contents of the panel via our MTL method.

Multitask Learning with Vision Transformers In its most conventional form, multitask learning predicts multiple outputs out of a shared encoder/representation for an input [50]. Prior works [19, 39, 40, 48, 49] follow this architecture to jointly learn multiple vision tasks using a CNN. Leveraging this encoder-decoder architecture, IPT [7] was the first transformer-based multitask network aiming to solve low-level vision tasks after fine-tuning a large pretrained network. This was followed by [25], which jointly addressed the tasks of object detection and semantic segmentation. Recently, [37] used a similar architecture for scene and action understanding and score prediction in videos. Following this, Hu et.al. [13] proposed a framework that tackles several language tasks but a single vision one. Differently, MulT [5] introduced a multitask transformer to handle multiple vision tasks. More complex vision transformer architectures have demonstrated that they outperform Convolutional Neural Network (CNN) based multitasking methods. However, neither do these existing methods generalize to the comics domains nor can they be trained to achieve predictions on comics imagery as they are fully-supervised networks.

With the development of image style transfer and its connection with domain adaptation, recently [3] adopted style transfer and adversarial training to estimate depth on comics. In essence, the style transfer [4] technique helps them to leverage models trained with large amounts of real-world ground-truth data. In this vein, we apply an unsupervised I2I translation method to minimize the domain disparity between comics and the real world.

2.1. Domain Adaptation via I2I Translation

The advent of I2I translation methods began with the invention of conditional GAN [24], which have been applied to a multitude of tasks, such as scene translation [15] and sketch-to-photo translation [43]. While conditional GANs yield impressive results, they require paired images during training. Unfortunately, in comics⟶real I2I translation scenario, such paired training data is lacking and expensive to collect. To overcome this, cycleGAN [52], with its cycle consistency loss between the source and target domains, is a possible solution for translating the comics images to real images, thereby producing consistent images. Nevertheless, neither conditional GANs, nor cycleGAN account for the multi-modality of comics⟶real I2I translation; in general, a single comics image can be translated to the real domain in many different, yet equally realistic ways. This is also due to the different artistic styles present in a single comics domain, which in turn, gives rise to intra-comics domain style variability. Addressing this issue of multi-modality, MUNIT [14] and DRIT [18] introduced solutions by learning a disentangled representation with a domain-invariant content space and a domain-specific attribute/style space. While effective, all the above-mentioned methods perform image-level translation, without considering the object instances. As such, they tend to yield less realistic results when translating complex scenes with many objects. This is also the task addressed by INIT [38] and DUNIT [4]. While INIT [38] proposed to define a style bank to translate the instances and the global image separately, DUNIT [4] proposed to unify the translation of the image and its instances, thus preserving the detailed content of object instances. We, therefore, use DUNIT [4] as our I2I translation model to translate the comics images to the real domain. Note that, we can also use a diffusion-based [30] translation method to achieve comics translations. Once translated, we leverage an MTL network trained with segmentation and depth annotations from real images, to ultimately, predict the segmentation and depth of comics images. To enforce the domain similarity between the comics and the real domains, we utilize a transformer-based domain discriminator. We now explain our method in detail.

3. Methodology

3.1. Problem Formulation and Overview

We aim to learn a cross-domain multitask mapping between two visual domains $C \subset \mathbb{R}^{H\times W\times 3}$ and $R \subset \mathbb{R}^{H\times W\times 3}$, where C is the comics domain and R is the real image domain. To this end, first, we employ the DUNIT model [4] to translate the comics image C to the real domain. This is an imperative step for achieving dense prediction on comics as we want to leverage the real-world annotations, thereby overcoming the lack of annotations in the comics domain. Moreover, simply applying an existing MTL model on these translated images does not help as comics images are highly disparate from real-world imagery. Hence, the models should be aware of the image semantic contents present in comics imagery, which is achieved by training the MTL model with the translated comics images.

Second, we use our MTL module on the translated image. Thus, the problem can be formulated as $TaskPred_c = f_{\mathcal{L}_{MTL}}(R(C))$, where $TaskPred_c$ are the multitask predictions for the image in the comics domain C, $R(C)$ is the comics⟶real translated image, and $f_{\mathcal{L}_{MTL}}(R(C))$ is the MTL module trained with the multitask loss

$$\mathcal{L}_{MTL} = w_{seg}\mathcal{L}_{CE} + w_{depth}\mathcal{L}_{rotate}\ . \tag{1}$$

on real images and applied to $R(C)$. Here, w_{seg} and w_{depth} are the weights of segmentation and depth tasks learned via GradNorm [8], respectively. $\mathcal{L}_{CE}$ is cross entropy loss for the segmentation task and $\mathcal{L}_{rotate}$ is the depth loss. The detailed architecture of our method is provided in Figure 2. We now explain the components of our network in more detail.

3.2. Image-to-Image Translation Module

Our method is built on the DUNIT [4] backbone which embeds the input images onto a shared style space and a domain-specific content space. As such, we use the same weight-sharing strategy as DUNIT for the two style encoders (E_c^s, E_r^s) and exploit the same loss terms. Here, $(E_c^{\cdot}, E_r^{\cdot})$ denote the encoders in the comics and real domains, respectively. They include:

- A content adversarial loss $\mathcal{L}_{adv}^{con}(E_c^{con}, E_r^{con}, D^{con})$ relying on a content discriminator D^{con} and the two content encoders (E_c^{con}, E_r^{con}), whose goal is to distinguish the content features of both the comics and real domains, respectively;

- Domain adversarial losses $\mathcal{L}_{adv}^{c}(E_r^{con}, E_c^{s}, G_c, D^c)$ and $\mathcal{L}_{adv}^{r}(E_c^{con}, E_c^{instcon}, E_r^{s}, G_r, D^r)$, one for each domain, with corresponding domain classifiers D^c and D^r, corresponding domain generators G_c and G_r and instance content encoder $E_c^{instcon}$;

- A cross-cycle consistency loss $\mathcal{L}_1^{cc}(G_c, G_r, E_c^{con}, E_c^{instcon}, E_r^{con}, E_c^{s}, E_r^{s})$ that exploits the disentangled content and style representations for cyclic reconstruction [45];

- Self-reconstruction losses $\mathcal{L}_{rec}^{c}(E_c^{con}, E_c^{instcon}, E_c^{s}, G_c)$, $\mathcal{L}_{rec}^{r}(E_r^{con}, E_r^{s}, G_r)$, one for each domain, ensuring that the generators can reconstruct samples from their own domain;

- KL losses for each domain $\mathcal{L}_{KL}^{c}(E_c^{s})$ and $\mathcal{L}_{KL}^{r}(E_r^{s})$ encouraging the distribution of the style representations to be close to a standard normal distribution;

- Latent regression losses $\mathcal{L}_{lat}^{c}(E_c^{con}, E_c^{instcon}, E_c^{s}, G_c)$ and $\mathcal{L}_{lat}^{r}(E_r^{con}, E_r^{s}, G_r)$, one for each domain, encouraging the mappings between the latent style representation and the image to be invertible;

- An instance consistency loss $\mathcal{L}_1^{ic}(P_{tl}^{ci}, P_{tl}^{ri}, P_{br}^{ci}, P_{br}^{r})$ encouraging the same object instances to be detected in the original comics image and in the corresponding image after translation, where $P_{(.)}^{(.)}$ are the bounding box top-left and bottom-right corner pixels for detected instances in the two domains.

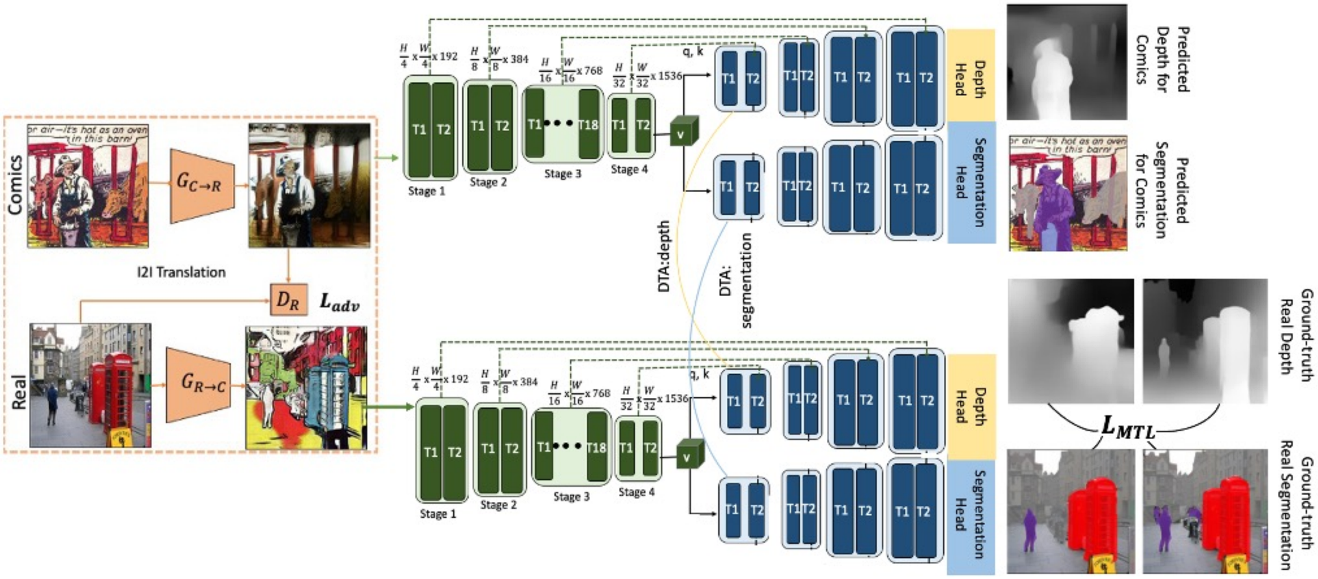
Comics
$G_{C\to R}$
I2I Translation
D_R
L_{adv}
Real
$G_{R\to C}$
$\frac{H}{4}\times\frac{W}{4}\times 192$
$\frac{H}{8}\times\frac{W}{8}\times 384$
$\frac{H}{16}\times\frac{W}{16}\times 768$
$\frac{H}{32}\times\frac{W}{32}\times 1536$
q, k
v
T1 T2
T1 ••• T18
Stage 1
Stage 2
Stage 3
Stage 4
DTA:depth
DTA: segmentation
Depth Head
Segmentation Head
Predicted Depth for Comics
Predicted Segmentation for Comics
Ground-truth Real Depth
Ground-truth Real Segmentation
L_{MTL}

Figure 2. **Detailed overview of our architecture for dense prediction on comics.** Our model comprises 1) an unsupervised image-to-image translation module to translate the comics images to the real domain and 2) a Swin transformer [21] based multitasking framework that performs segmentation and depth estimation, simultaneously while bridging the domain gap using a Domain Transferable Attention (DTA). The encoder module (in **green**) embeds a shared representation of the input image, which is then decoded by the transformer decoders (in **blue**) for the respective tasks. Note that the transformer decoders have the same architecture but different task heads. The MTL model is trained using a weighted loss [8] of all the tasks involved.

During training, the I2I translation module is trained separately and then we apply our MTL module.

3.3. MTL Module

Our MTL module follows the principle of a transformer encoder-decoder architecture [44]. It consists of a transformer-based encoder to map the input image to a latent representation shared by the tasks, followed by transformer decoders with task-specific heads producing the predictions for the respective tasks. Figure 2 shows an overview of our MTL framework. For our transformer-based encoder, we use a pyramidal backbone, named the Swin Transformer [21] to embed the visual features into a list of hidden states that incorporates global contextual information. We then apply the transformer decoders to progressively decode and upsample the tokenized maps from the encoded image. Finally, the representation from the transformer decoder is passed to a task-specific head, such as a simple two-layer classifier (in the case of segmentation), which outputs the final predictions. Given the simplicity of our network, it can be extended easily to more tasks. The following sections describe the details of each component of our network.

3.3.1 Encoder Module

For the encoder, we adopt Swin-B [21], which applies stacked transformers to features of gradually decreasing resolution in a pyramidal manner, hence producing hierarchical multi-scale encoded features, as shown in Figure 2. In particular, following the ResNet [12] structure and design rules, four stages are defined in succession: each of them contains a patch embedding step, which reduces the spatial resolution and increases the channel dimension, and a columnar sequence of transformer blocks. This approach halves the resolution and doubles the channel dimension at every intermediate stage, matching the behavior of typical fully-convolutional backbones and producing a feature pyramid (with output sizes of 1/4, 1/8, 1/16, 1/32 of the original resolution) compatible with most previous architectures for vision tasks.

Out of a total of $N = 24$ transformer encoders, 2 blocks are in the first, second, and fourth stages, and 18 are in the third stage. In each block, self-attention is repeated according to the number of heads used and depending on the stage of the encoding process. This is done to match the increase in the channel dimensions, where the dimensions are $M = \{6, 12, 24, 48\}$ in the first, second, third, and fourth stages, respectively. However, the high resolution in the first two stages does not allow the use of global self-attention, due to its quadratic complexity with respect to the token sequence length. To solve this issue, in all stages, the tokens, that are reshaped in a 2D representation, are divided into non-overlapping square windows of size $h = w = 7$, and the intra-window self-attention is independently computed for each of them. This means that each token attends to only the tokens in its own window, both as a query and as a key/value. A possible downside of this approach could be that the restriction to fixed local windows completely stops any type of global or long-range interaction. The adopted solution is to alternate regular window partitioning with another non-overlapping partitioning in which the windows are shifted by half their size, $\lfloor h/2 \rfloor = \lfloor w/2 \rfloor = 3$, both in the height and width dimensions. This has the effect of gradually increasing the virtual receptive field of the subsequent attention computations.

3.3.2 Decoder Module

Inspired by the two CNN-based decoders proposed in [51], we use corresponding conceptually similar transformer-based versions. The general idea is to replace convolutional layers with windowed transformer blocks. Specifically, our decoder architecture consists of four stages, each containing a sequence of 2 transformer blocks for a total of 8. In each stage, the two sequential transformer blocks allow us to leverage inter-window connectivity by alternating regular and shifted window configurations as in the encoder. Between consecutive stages, we use an upsampling layer to double the spatial resolution and half the channel dimension; we, therefore, adjust the number of attention heads accordingly to 48, 24, 12, 6, in the first, second, third, and fourth stage, respectively. The spatial/channel shape of the resulting feature maps matches the outputs of the encoder stages, which are delivered to the corresponding decoder stages by skip connections. This yields an hourglass structure with mirrored encoder-decoder communication: the lower-resolution stages of the decoder are guided by the higher-level deeper encoded features and the higher-resolution stages of the decoder are guided by the lower-level shallower encoded features, allowing to gradually recover information in a coarse to fine manner and to exploit the different semantic levels where they are more relevant.

To perform multitask prediction, we share the encoder across all tasks and use task-specific decoders with the same architecture but different parameter values. We then simply append task-specific heads to the decoder. For instance, a model jointly trained for semantic segmentation and depth prediction will have two task-specific heads: one predicting K channels followed by a softmax for semantic segmentation and one predicting a single channel followed by a sigmoid for depth estimation.

3.3.3 Domain Transferable Attention

For effective knowledge transfer between the two domains, it is essential to focus on both transferable and discriminative features which can be leveraged from 1) the fine-grained feature tokens of Swin transformer and 2) the attention weights in the Swin transformer that convey discriminative information of the tokens between the domains. While the self-attention weights in Swin could be employed to discriminate between the domains, one major hurdle here is, the feature tokens do not transfer across domains. We, therefore, use a modified Domain Transferable Attention (DTA) mechanism [47] that integrates the feature tokens of the real domain into the translated comics⟶real domain stream. To this end, we employ a token-level domain discriminator D_{token} that matches cross-domain local features by optimizing:

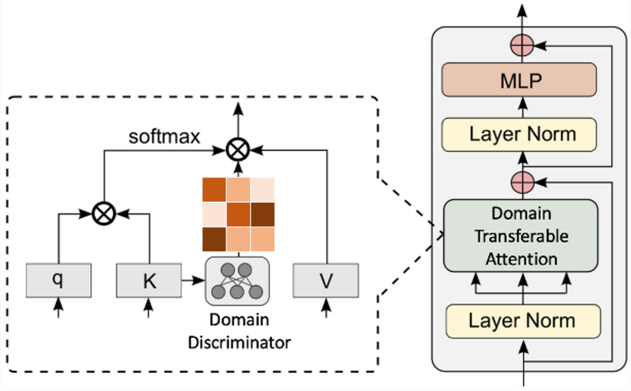

Figure 3. Overview of the **domain transferable attention** mechanism.

$$\mathcal{L}_{token}\left(x^{c}, x^{r}\right) = -\frac{1}{n}\sum_{x_1 \in C}\sum_{i=1}^{n}\mathcal{L}_{MTL}\left(D_{token}\left(G_f\left(x_i^{R\to C}\right)\right)\right). \quad (2)$$

where n is the number of feature tokens, and $D_{token}(.)$ is the probability of the feature token belonging to the comics domain. During adversarial learning, D_{token} tries to assign 1 for a feature token belonging to the comics domain and 0 for those belonging to the real domain, while G_f takes the real domain feature tokens and produces fake comics tokens, denoted by $x_i^{R\longrightarrow C}$. Conceptually, a feature token that can easily deceive D_{token} is more transferable across domains and should be given a higher transferability, i.e., a higher value of D_{token} for a feature token in real domain implies higher transferability of that token to the comics domain. We, therefore, measure the transferability of a feature token in the real domain to that of comics by $T(.) = H\left(D_{token}(.)\right) \in [0,1]$, where $H(.)$ is the standard entropy function. This procedure allows us to leverage the tokens from the real domain that are learned via full supervision, thanks to the available real-world ground-truth annotations. We then inject the transferable tokens from the real domain into the self-attention weights computed in the comics domain (top stream in Figure 2) as follows:

$$\mathrm{DTA}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{q}^{T}}{\sqrt{d}}\right) \odot \left[1; T\left(\mathbf{K}_{\text{token}}\right)\right] \mathbf{V}\,. \quad (3)$$

where $\mathbf{q}$ is the query of the feature token, $\mathbf{K}_{\text{token}}$ is the key of the transferable feature token from the real domain to the comics domain, $\odot$ is Hadamard product, and [;] is concatenation operation. Obviously, softmax $\left(\frac{\mathrm{qK}^T}{\sqrt{d}}\right)$ and $[1;T\left(\mathbf{K}_{\text{token}}\right)]$ indicate the discrimination (semantic importance) and the transferability of each token, respectively. To jointly attend to the transferability of different tasks and of different locations, we thus define DTA as:

$$\begin{aligned} \operatorname{DTA}(\mathbf{q}, \mathbf{K}, \mathbf{V}) &= \operatorname{Concat}\left(\text{head}_1, \ldots, \text{ head }_k\right) \mathbf{W}^o \\ \text{where head } &= \operatorname{SA}\left(\mathbf{q} \mathbf{W}_i^q, \mathbf{K} \mathbf{W}_i^K, \mathbf{V} \mathbf{W}_i^V\right) \end{aligned} \tag{4}$$

Note that SA is the self-attention of the transformer. This is followed by the residual and LayerNorm operation. Subsequently, an MLP and another residual operation is carried out as follows:

$$\begin{aligned} \hat{z}^l &= \operatorname{DTA}\left(\operatorname{LN}\left(\mathbf{z}^{l-1}\right)\right)+\mathbf{z}^{l-1} \\ \mathbf{z}^l &= \operatorname{MLP}\left(\operatorname{LN}\left(\hat{\mathbf{z}}^l\right)\right)+\hat{\mathbf{z}}^l \end{aligned} \tag{5}$$

By applying the above procedure in our MTL module, we can enforce the domains to be transferable. This, in turn, allows us to achieve dense prediction in the comics domain by leveraging supervision from real-world annotations.

Note that DTA differs from the co-attention introduced in prior works [5, 6], wherein both cases, the attention is computed based on a specific *task*. By contrast, we learn a domain transferable attention between different *domains*.

Task Heads and Loss. The feature maps from the transformer decoder modules are input to different task-specific heads to make subsequent dense predictions in the comics domain, denoted by $TaskPred_c$. Each class head includes a single linear layer to output a $H \times W \times 1$ map, where H, W are the input image dimensions. We employ a weighted sum $\mathcal{L}_{MTL}$, as stated in Equation 1, to train the MTL network, where the losses are calculated between the ground truth and final predictions for each task in the real domain. In particular, we use cross-entropy for segmentation and rotate loss [48] for depth. Note that we employ these losses to maintain consistency with the baselines [39, 48].

4. Experiments and Results

To validate our method, we conduct experiments on the following datasets.

4.1. Datasets

The main dataset used for this work is the DCM dataset that comprises 772 full-page images with multiple comics panel images within. We extract 4470 single panel images from these full-page images using the panel annotations. Note that the panel annotations do not contain semantic or depth information. We thus, use these DCM panel images to translate them to real ones using [4]. Following this, we apply our MTL model on the translated images. We evaluate the performance on the DCM validation set that contains dense depth annotations for 300 DCM comics images and was introduced by [3]. For evaluating on semantic segmentation, we use the OpenCV CVAT interface [11], leveraging the semantic labels of MS-Coco. We further test our method for dense predictions on a novel comics test dataset of Spirou [35] and Tintin.

4.2. Evaluation Metrics

To evaluate our method, we evaluate the following two standard performance metrics, for the tasks of segmentation and depth, respectively. These metrics were reported for consistency with the baselines [5, 21, 36, 48].

Semantic segmentation uses *mIoU* as the average of the per-class Intersection over Union (%) between the ground-truth segmentation and predicted map.

Depth uses the Root Mean Square Error *(RMSE)* computed between the depth label and the predicted depth map, where the RMSE metric is reported in meters over the evaluated set of images.

4.3. Training Details

We use a pretrained DUNIT module [4] to first, translate the comics images to real ones. This was done for all the baselines as well. We train each baseline as per their best configurations for the tasks of segmentation and depth, mentioned in their respective works [5, 21, 36, 48].

We, then, train our MTL method on semantic segmentation and depth estimation. In our implementation, we train with a batch size of 8 on 2 Nvidia A100-40GB GPU, using PyTorch. We use the weighted Adam optimizer [22] with a learning rate of 5e-5 and the warm-up cosine learning rate schedule (using 2000 warm-up iterations). The optimizer updates the model parameters based on gradients from the task losses.

4.4. Baselines

We compare our MTL model with the following state-of-the-art baselines.

Baseline UNet [36] (for single-task learning) constitutes our CNN-based baseline. We use it as a reference for all the multitask models.

Baseline Swin transformer [21] (for single-task learning) constitutes the single task transformer baseline. It is almost identical to our MTL model, except it does not include DTA, and is trained with only one dedicated task. We use it to evaluate the benefits of our multitask learning strategy

Consistency [48] presents a general and data-driven framework for augmenting standard supervised learning with cross-task consistency. Based on a CNN backbone, it is inspired by Taskonomy [49] but adds a consistency constraint to learn multiple tasks jointly.

MulT [5] comprises a Swin transformer-based multitasking network with one shared encoder and multiple decoders each dedicated to a task. This baseline further identifies if tasks are interdependent, such that a shared representation can give comparable performance across multiple tasks, without explicitly adding task constraints. For MulT, we use depth as the reference task because they report the best performance with depth as a reference when jointly trained with segmentation.

Note that, we do not compare with existing works in comics analysis as none of them perform dense predictions such as segmentation or depth estimation. All the multitask baselines were trained using their best model configurations for segmentation and depth, as in [5, 48], respectively. All the methods were evaluated on the translated comics image using [4] for a fair comparison.

4.5. Quantitative Results

Table 1 shows the comparative performance of all the evaluated baselines on the DCM validation set. Our model outperforms the multitask CNN-based baseline [48] as well as the multitask Swin transformer baseline [21] when the MTL models are jointly trained on segmentation and depth, denoted as the ’S-D’ setting. We also considerably outperform the single-task CNN baseline [36] and the single-task Swin baseline [21] that are trained on isolated tasks. We, therefore, report improvements in both segmentation and depth in comparison to all the baselines. Note that Tintin and Spirou are test datasets and we provide only qualitative comparisons on them.

Quantitative results on DCM [28]			***’S-D’***	
	Methods	MTL	SemSeg mIoU%↑	Depth RMSE↓
CNN	UNet [36]		23.64	1.033
	Cross-task Consistency [48]	✓	24.72	0.999
Transformer	1-task Swin [21]		28.95	0.958
	MulT [5]	✓	<u>30.00</u>	<u>0.955</u>
	Our	✓	**33.65**	**0.909**

Table 1. **Quantitative results for multitasking on DCM validation set** [28]. Our model outperforms both the single-task [21, 36] and multitiask [5, 48] baselines. Bold and underlined values show the best and second-best results, respectively.

4.6. Qualitative Results

Qualitatively, we show, in Figures 4 and 5, the segmentation and depth results of the best-performing models from Table 1. The models are applied to 1) the translated DCM validation images, 2) the translated Spirou test images, and 3) the translated Tintin test images, respectively. For both segmentation and depth, we show that our method clearly outperforms the baselines, including MulT which is a transformer-based handcrafted model for MTL. Specifically, in Figure 4, for the Spirou test image, our method is the only approach that is able to segment the dog (shown in orange segmentation mask) whereas all the baselines fail to segment it. Also, our method achieves more accurate segmentation than both the single-task as well as the multitask baselines for the ’person’ category in DCM, Spirou, as well as Tintin images (shown in purple segmentation masks). We also show that our method achieves better foreground versus background separation in the depth estimates in Figure 5, across all the images. Particularly, for the DCM image in Figure 5, our method accurately predicts the depth plane of the people whereas the baseline single-task Swin and the multitasking method (MulT) fail to do so.

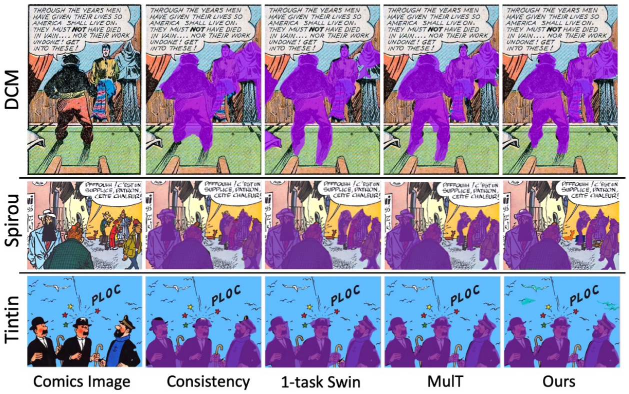
DCM
Spirou
Tintin
THROUGH THE YEARS MEN HAVE GIVEN THEIR LIVES SO AMERICA SHALL LIVE ON. THEY MUST NOT HAVE DIED IN VAIN.... NOR THEIR WORK UNDONE! GET INTO THESE!
PFFFOUHH ! C'EST UN SUPPLICE, PATRON, CETTE CHALEUR!
PLOC
Comics Image
Consistency
1-task Swin
MulT
Ours

Figure 4. **Qualitative comparison of semantic segmentation** on a DCM validation [28] image, a Spirou test image, and a Tintin test image, respectively. We show, from left to right, the input image in the comics domain, the results using the multitask CNN-based model (Consistency) [48], the single-task Swin transformer-based segmentation model (1-task Swin) [21], the multitask Swin transformer-based model (MulT) [5], and our model, respectively. Best viewed in color.

Module	SemSeg mIoU%↑	Depth RMSE↓
I2I+MTL	<u>29.36</u>	<u>0.958</u>
I2I+MTL+DTA **(Our)**	**33.65**	**0.909**

Table 2. **Ablation Study on DCM validation set** [28]. We add the modules of our network one by one to study their effect on task performance. The DTA mechanism significantly benefits the performance by transferring the feature tokens between the real and the comics domain. Bold and underlined values show the best and second-best results, respectively.

4.7. Ablation Study

In Table 2, we study the effect of the different components of our method. We do not isolate the I2I module as it is a necessary pre-processing step to acquire translated images. Without the translation, all the methods fail to infer comics images. Explicitly, we study the effect of the DTA mechanism on our MTL module and find that it significantly benefits dense predictions. Without the DTA mechanism,

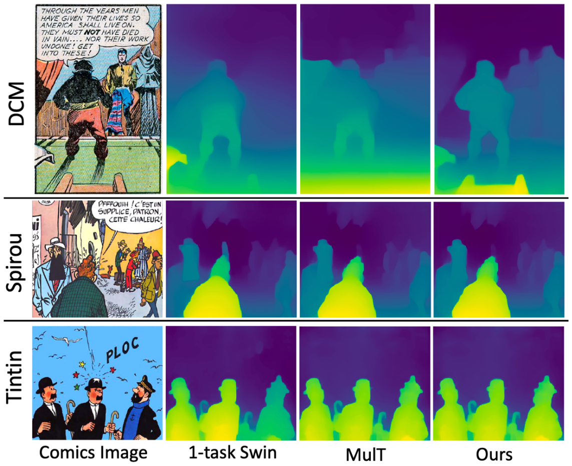
DCM
Spirou
Tintin
THROUGH THE YEARS MEN HAVE GIVEN THEIR LIVES SO AMERICA SHALL LIVE ON... THEY MUST NOT HAVE DIED IN VAIN.... NOR THEIR WORK UNDONE! GET INTO THESE!
PFFFOUH ! C'EST UN SUPPLICE, PATRON. CETTE CHALEUR!
PLOC
Comics Image
1-task Swin
MulT
Ours

Figure 5. **Qualitative comparison of depth estimation** on a DCM validation [28] image, a Spirou test image, and a Tintin test image. We show, from left to right, the input image in the comics domain, the results using the single-task Swin transformer-based depth model (1-task Swin) [21], the multitask Swin transformer-based model (MulT) [5], and our model, respectively. Best viewed in color.

our performance is comparable to that of the handcrafted transformer-based MTL approach by MulT [5].

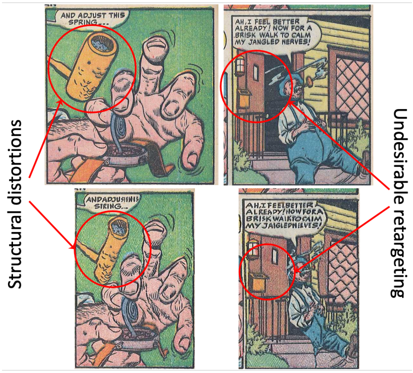
Structural distortions
Undesirable retargeting
AND ADJUST THIS SPRING...
AH, I FEEL BETTER ALREADY! NOW FOR A BRISK WALK TO CALM MY JANGLED NERVES!

Figure 6. **Limitations of the seam-carving method** [2] on comics images.

4.8. Application of our MTL Method to Retargeting

Our MTL method can be applied to retarget comics, thereby reconfiguring them to different digital media. Particularly, existing deep retargeting methods [17, 42] utilize implicit semantic and depth cues to retarget semantic units accurately on the correct depth plane. However, explicitly leveraging dense prediction features to guide the learning of the retargeting method has not yet been explored. One may ask: why should one use the dense prediction cues in a deep learning framework when one can simply apply energy computations as done in the widely known seam-carving [2] method? The reason is that seam-carving, due to its non-differentiability, gives rise to structural distortions and undesirable retargeting on comics images as shown in Figure 6. Therefore, we turn to differentiable deep networks and employ additional cues from our dense task predictions to guide the retargeting methodology. In particular, we can leverage our MTL model to aid an off-the-shelf deep learning retargeting method like [17] to retarget comics images. The results are shown in Figure 7 where the uninformative areas of the images are removed while preserving their contents and narrative. This is one such example of retargeting comics panels to a given device size. Applying our method on a large-scale and across different media will, ultimately, help authors to reconfigure their works to diverse media in an automated manner.

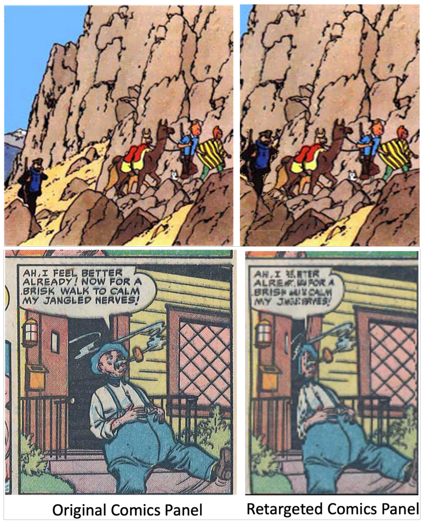
AH, I FEEL BETTER ALREADY! NOW FOR A BRISK WALK TO CALM MY JANGLED NERVES!
Original Comics Panel
Retargeted Comics Panel

Figure 7. **Applying our MTL model to retarget comics.** We integrate the dense predictions from our MTL method to guide the retargeting algorithm [17] to achieve results on Tintin (Top) and DCM (Bottom) test images.

5. Conclusion

In summary, while the state-of-the-art in comics analysis remains limited to detecting panels, text, and bounding boxes for specific characters, we achieve detailed segmentation of the generic comics elements as well as infer notions of 3D from them. Benefiting from these dense predictions, this work can have different applications, such as comics scene understanding, and retargeting the comics panels. We demonstrate the applicability of our developed method by integrating it with an off-the-shelf retargeting algorithm, thereby automatically reconfiguring comics panels. This will open up possibilities to help comics authors to diffuse their work across different publication channels, thus benefiting the comics industry.

Acknowledgement. This work was supported in part by the Swiss National Science Foundation via the Sinergia grant CRSII5−180359.

References

[1] Kohei Arai and Herman Tolle. Method for real time text extraction of digital manga comic. *International Journal of Image Processing (IJIP)*, 4(6):669–676, 2011. 1, 2

[2] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, 26(3):10–es, July 2007. 8

[3] Deblina Bhattacharjee, Martin Everaert, Mathieu Salzmann, and Sabine Süsstrunk. Estimating image depth in the comics domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2070–2079, January 2022. 1, 2, 6

[4] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4787–4796, 2020. 2, 3, 6, 7

[5] Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12031–12041, June 2022. 2, 6, 7, 8

[6] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021. 6

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *arXiv: 2012.00364, cs.CV*, 2021. 2

[8] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 794–803, 2018. 3, 4

[9] Wei-Ta Chu and Wei-Chung Cheng. Manga-specific features and latent style model for manga style analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1332–1336. IEEE, 2016. 2

[10] Wei-Ta Chu and Wei-Wei Li. Manga facenet: Face detection in manga based on deep neural network. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 412–415, 2017. 1, 2

[11] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), 9 2022. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015. 4

[13] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv: 2102.10772, cs.CV*, 2021. 2

[14] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *ECCV*, abs/1804.04732, 2018. 3

[15] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36:1–14, 07 2017. 2

[16] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost Van De Weijer, Andrew D Bagdanov, Maria Vanrell, and Antonio M Lopez. Color attributes for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3306–3313, 2012. 2

[17] Gil Laufer. *Video Retargeting using Vision Transformers: Utilizing deep learning for video aspect ratio change*. PhD thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 06 2022. 8

[18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 2018. 3

[19] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. *CoRR*, abs/1803.10704, 2018. 2

[20] Xicheng Liu, Yongtao Wang, and Zhi Tang. A clump splitting based method to localize speech balloons in comics. In *Proceedings of the 13th IEEE International Conference on Document Analysis and Recognition (ICDAR)*, pages 901–905, 2015. 1

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv: 2103.14030, cs.CV*, 2021. 2, 4, 6, 7, 8

[22] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *Openreview*, 2018. 6

[23] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 2

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 2

[25] Eslam Mohamed and Ahmed El-Sallab. Spatio-temporal multi-task learning transformer for joint moving object detection and segmentation. *arXiv: 2106.11401, cs.CV*, 2021. 2

[26] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Comic characters detection using deep learning. In *Proceedings of the 14th International Conference on Document Analysis and Recognition (ICDAR)*, volume 3, pages 41–46, 2017. 1, 2

[27] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. *Journal of Imaging*, 4(7):89, 2018. 2

[28] Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. *Journal of Imaging*, 4(7), 2018. 7, 8

[29] Xiaoran Qin, Yafeng Zhou, Zheqi He, Yongtao Wang, and Zhi Tang. A faster r-cnn based method for comic characters face detection. In *Proceedings of the 14th IEEE International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1074–1080, 2017. 1, 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 3

[31] Christophe Rigaud. Segmentation and indexation of complex objects in comic book images. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 14, 12 2014. 1, 2

[32] Christophe Rigaud, Jean-Christophe Burie, and Jean-Marc Ogier. Text-independent speech balloon segmentation for comics and manga. In *Proceedings of the 11th International Workshop on Graphic Recognition: Current Trends and Challenges*, pages 133–147, 2017. 1

[33] Christophe Rigaud, Nam Le Thanh, J-C Burie, J-M Ogier, Motoi Iwata, Eiki Imazu, and Koichi Kise. Speech balloon and speaker association for comics and manga understanding. In *Proceedings of the 13th IEEE International Conference on Document Analysis and Recognition (ICDAR)*, pages 351–355, 2015. 1

[34] Christophe Rigaud, Norbert Tsopze, Jean-Christophe Burie, and Jean-Marc Ogier. Robust frame and text extraction from comic books. In *Proceedings of the International Workshop on Graphics Recognition: New Trends and Challenges*, pages 129–138. Springer, 2013. 1, 2

[35] Rob-Vel. Spirou, 1947. 6

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 6, 7

[37] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Video multitask transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshop (ICCVW)*, pages 1553–1561, 2019. 2

[38] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S. Huang. Towards instance-level image-to-image translation. *CoRR*, abs/1905.01744, 2019. 3

[39] Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? *arXiv: 1905.07553, cs.CV*, 2019. 2, 6

[40] Gjorgji Strezoski, Nanne van Noord, and Marcel Worring. Many task learning with task routing. *arXiv:1903.12117*, 2019. 2

[41] Weihan Sun, Jean-Christophe Burie, Jean-Marc Ogier, and Koichi Kise. Specific comic character detection using local feature matching. In *Proceedings of the 12th IEEE International Conference on Document Analysis and Recognition*, pages 275–279. IEEE, 2013. 2

[42] Weimin Tan, Bo Yan, Chuming Lin, and Xuejing Niu. Cycle-ir: Deep cyclic image retargeting. *IEEE Transactions on Multimedia*, 22(7):1730–1743, 2019. 8

[43] Radim Tyleček and Radim Šára. Spatial pattern templates for recognition of objects with regular structure. In *Proceedings of the German Conference on Pattern Recognition*, pages 364–374, 2013. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 4

[45] Chengjia Wang, Gillian Macnaught, Giorgos Papanastasiou, Tom MacGillivray, and David E. Newby. Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks. *CoRR*, abs/1808.03944, 2018. 3

[46] Masashi Yamada, Rahmat Budiarto, Mamoru Endo, and Shinya Miyazaki. Comic image decomposition for reading comics on cellular phones. *IEICE Transactions on Information and Systems*, 87(6):1370–1376, 2004. 1, 2

[47] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2108.05988*, 2021. 5

[48] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11197–11206, 2020. 2, 6, 7

[49] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[50] Yu Zhang and Qiang Yang. A survey on multi-task learning. *arXiv:1707.08114, cs.LG*, 2021. 2

[51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2